Getting the Most out of a Limited Sample Size Field Test: Experiences from the National Survey on Drug Use and Health

Jonaki Bose[1], Dicy Painter[1], Doug Currivan[2], Larry Kroutil[2], & Kevin Wang[2]

[1]Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 1 Choke Cherry Rd, Rockville MD 20857
[2]RTI International; Research Triangle Park, NC 27709

Overview

This paper uses experiences from the survey redesign process of the National Survey on Drug Use and Health (NSDUH), including a field test. The paper describes ways in which even a limited-sample field test can inform decision making regarding the redesign or changing of a survey.

The NSDUH is the federal government's primary source of information on the nature and extent of substance use in the United States. Conducted since 1971, the survey collects data from a representative sample of about 67,500 persons in the United States at their place of residence. The respondent universe is the civilian, noninstitutionalized population aged 12 years old or older in the United States. Persons who are outside of this population include active-duty military personnel, persons with no fixed household address (e.g., homeless or transient persons not in shelters), and residents of institutional group quarters, such as prisons and long-term hospitals. Young people are oversampled, with one-third of the sample in each state being allocated to each of the following three age groups: 12-17, 18-25, and 26 and older. At each sampled address, interviewers conduct a 5-minute screening procedure using a handheld computer to list all household members and their basic demographic data. To obtain the target sample sizes, a preprogrammed selection algorithm selects zero to two sample person(s) aged 12 or older, depending on the composition of the household.

The main survey data are collected through face-to-face computer-assisted interviewing (CAI), including audio computer-assisted self-interviewing (ACASI), on a laptop computer. The interviews average about an hour. Each respondent who completes a full interview is given a $30 cash incentive. The survey is sponsored by the Substance Abuse and Mental Health Services (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ) and the data are collected under contract by RTI International.

The questionnaire contains questions pertaining to the use of tobacco, alcohol, and illicit drugs, as well as injection drug use, perceived risks of substance use, substance dependence and abuse, arrests, treatment for substance use problems, pregnancy, health conditions, health care utilization, and items on mental health. Mental health items contained in the NSDUH interview include questions about use of mental health services, short scales that measure psychological distress (the Kessler-6 [K6] scale) and functional impairment due to psychological distress (adapted World Health Organization Disability Assessment Schedule [WHODAS]), items on major depressive episode and items on serious thoughts of suicide and suicidal behavior.

Most of the questions are administered through ACASI. ACASI is designed to provide the respondent with a highly private and confidential mode for responding to questions in order to increase accurate reporting of illicit drug use and other sensitive behaviors. Less sensitive items such as demographic items and background items on income and health insurance are administered by interviewers using computer-assisted personal interviewing (CAPI).

In addition to the mode of administration, questions on the survey can be distinguished by whether or not the questions can be modified from one year to the next. Core questions, such as key demographic characteristics and drug use prevalence questions, have been designed to stay relatively constant from one year to the next to permit measurement of trends in drug use across time. In contrast, the content of noncore questions can change across years to measure new or developing topics of interest or to rotate certain topics in or out of the interview.

The NSDUH Redesign Process

In order to continue producing data that accurately reflect current conditions, the NSDUH must be updated periodically to address changing substance use and mental health issues. A redesign can be used to update the questionnaire and sample design to meet current data needs and measure new phenomena. It also allows SAMHSA to determine whether generating state and substate estimates continues to be a priority. In theory, a redesign may allow for implementation of better data collection, processing and estimation methods. A redesign may also present cost-saving measures that would help bring survey costs closer to expected budget levels. More specifically, the NSDUH redesign allows for the updating of outdated prescription drug questions and addresses the increased need for data on mental health, screening for substance use, military families and other subgroups of interest.

Because the NSDUH is designed to measure trends across time, there is a trade-off between making improvements to the survey and the potential for breaking trends. Thus, any redesign process has to assume that a redesign will mark the start of a new baseline for trend data or else face the additional challenge of improving survey processes without affecting the data that are being collected and the estimates that are produced.

CBHSQ is planning to implement changes related to a partial NSDUH redesign, principally in 2014 and 2015. These changes include a modified sample design in 2014 with new state and age sample allocations and the elimination of the half open interval (HOI) procedures for addressing noncoverage. A limited update to the interview questionnaire will be made in 2015. Additionally, in 2013, questions related to serving in the military reserves, medical marijuana use, respondent height and weight, use of primary care services in the past year, and discussions with a doctor regarding substance use in the past year were also implemented.

The new sample design will allow for continued national, state, and substate-level estimation that are comparable with estimation procedures from previous surveys. The new sample design's improved efficiency will result in significant cost savings. The primary change to the questionnaire will be to update the prescription drug questions. Changes to the prescription drug questions will include a new question structure that focuses on any use and misuse (also known as nonmedical use) of specific prescription drugs in the past 12 months, a revised definition of misuse that includes overuse of prescribed medication, and a focus on currently available prescription drugs. Other planned changes to the questionnaire include a revised health module that contains new questions about medical conditions, a separate methamphetamine module, new items on sexual orientation, military families and disability, and other updates. Other changes to the survey include new contact materials and equipment, change in the sponsor presented to the respondent (U.S. Public Health Service to the U.S. Department of Health and Human Services), use of on-screen pictures of prescription drugs and calendars instead of physical show cards, and the move of some items from CAPI to ACASI. These changes will seek to achieve three main goals: (1) to revise the questionnaire to address changing policy and research data needs, (2) to modify the survey methodology to improve the quality of estimates and the efficiency of data collection and processing, and (3) where appropriate, to maintain trends in core substance use estimates[1] across survey years.

The 2012 NSDUH field test was meant to test the revisions to the questionnaire and protocols. Additionally, the field test provided the first opportunity to see how the new items performed that had been added to the questionnaire in 2013. However, the field test was only one part of the approach being used by CBHSQ to make decisions regarding the redesign. Prior to the field test, CBHSQ assessed the data needs of current and potential data users, including groups such as states, national associations, other federal agencies, agencies within the U.S. Department of Health and Human Services (HHS), and other centers within SAMHSA. There also were a number of methodological studies that examined improvements to the contact materials, questionnaire structure, new questions, weighting, imputation, small area estimation methods, and data collection methods, including the use of alternate sampling frames. Various design alternatives were explored in terms of cost and impact on data quality and analytic capability. These studies utilized methods such as cognitive testing, analysis of existing data, and literature reviews. Based on this work and the requirements outline by management, options were presented to CBHSQ and SAMHSA leadership. Once an option was identified the 2012 field test was implemented. Following the 2012 field test, there was an additional dress rehearsal in 2013 to allow further testing of the procedures, materials, and questionnaire, including interviews in

[1] Drugs defined as core substance use items in NSDUH include tobacco, alcohol, marijuana, cocaine, crack cocaine, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives.

Spanish. As stated previously, the new sample design will be implemented in 2014, followed by the partial questionnaire redesign in January of 2015.

One of the challenges was that there were less than 12 months between the end of the field test data collection by the end of October 2012 and the start of the dress rehearsal in the beginning of September 2013. During that time, the field test data needed to be edited, imputed, and weighted; recoded variables needed to be created, tables had to be run and interpreted; and decisions for the dress rehearsal based on the field test results had to be made. Based on these decisions, changes also had to be made to the CAI instrument, training materials had to be developed, and OMB clearance needed to be obtained in advance of the start of data collection for the dress rehearsal.

The NSDUH 2012 Field Test

Prior to the implementation of the partial redesign, there are a number of questions for which we ideally would have liked to have answers. These included the following:

- *With the new design, can trends still be maintained for core items such as tobacco, alcohol, and marijuana?*
  One of the restrictions on the redesign was that the changes should not affect the ability to measure trends for the core drug sections of the interview that were not being updated.

- *How did moving items from CAPI to ACASI affect estimates of health insurance coverage, sources of income, personal and family income, and employment?*
  Since these items are key analytic variables, especially with the implementation of the Affordable Care Act (ACA), it was important to know if the change in modes affected estimates for these measures, producing data that are not comparable with those collected under the previous design.

- *What did estimates from new questions look like, particularly from the new prescription drug module?*
  The field test was our first chance to see what the new estimates might look like. Given the high level of interest in these items, we wanted to anticipate how the new estimates might compare to previous NSDUH estimates and other known estimates prior to fielding the partially redesigned questionnaire in 2015.

- *How long was the redesigned survey overall, and for specific modules (e.g., prescription drugs), and for particular subgroups of respondents (e.g., recent users of multiple substances or older respondents)?*
  The administration time for the survey is designed to be approximately 1 hour on average. However, we know that the survey typically takes longer for certain subgroups such as those who are heavy substance users. Older respondents also typically take longer to complete the interview. It was possible that the new prescription drug module would affect the administration times for these subgroups because it includes items on both use and misuse of prescription drugs in the past 12 months. The use of prescription drugs—regardless of misuse—typically increases with age. The emphasis on misuse of specific prescription drugs in the past 12 months could increase the administration time for respondents who misused multiple prescription drugs in that period.

- *How did the new equipment perform in the field?*
  Based on the results of the field tests, decisions would need to be made about the equipment to be provided to the more than 500 interviewers who will be collecting data in 2015.

- *How did the updated contact materials perform in the field*
  QFT interviewers answered debriefing items asking whether sample members' recalled receiving the lead letter and whether they made any comments on the updated lead letter or the Q&A brochure.

- *What was the interviewer and respondent experience like?*
  One of the goals of the redesign was to decrease the burden on interviewers by taking steps such as moving items from CAPI to ACASI and dropping the HOI. In addition, with the changes to the some of the survey components, we wanted to understand if the respondent experience was improving or changing negatively.

Even if methodological studies and cognitive tests were conducted, field testing allows changes to be tested under actual field conditions. Field testing with a probability sample also provides the opportunity to anticipate the effects of changes on factors such as estimates, trends, data quality, and respondent burden in advance of full-scale implementation. Field test results provide the opportunity for deciding whether to make certain changes or choosing among available options in preparation for the actual implementation.

However, if the field test design has limitations, then making inferences and decisions becomes challenging. Ideally, a field test should mimic all essential survey conditions as in the main survey. These include factors such as the content of contact materials, interviewer characteristics, field staff working conditions and level of effort, and response rates. Generally, if the aim of conducting a field test is to understand the effect of particular changes, then its design should allow the independent effects of these changes to be isolated and assessed. This is typically achieved through split sample survey designs. The sample sizes of a field test should correspond to the analytic goals, and very importantly, sufficient time must be allocated to analyze data and make informed decisions.

The NSDUH 2012 field test had a final sample size of 2,044 respondents. Even though it would have been useful to understand the effects on estimates that could be attributed to specific changes, there was no split sample. In addition, Alaska and Hawaii were not included in the field test. Unlike the regular NSDUH interview, which respondents can complete in either English or Spanish, the field test interview was available only in English. Also, experienced field staff were disproportionately used for data collection. Although the new age allocation was used, other elements of the new sample design that will be implemented in 2014 were not used. Due to the small sample size and time limitations, modified editing, imputation and weighting procedures were used for both the field test and main survey comparison data. Even though a sample size of approximately 2,000 cases is larger than many field tests, it was still small relative to an ideal sample size with sufficient statistical power to detect important differences. Consequently, the field test sample size constrained the kinds of inferences and decisions that realistically could be made. However, decisions still needed to be made regardless of the results of field testing. Therefore we attempted to make maximum use of all the resources that were available to us.

Maximizing the Utility of the 2012 Field Test

As described previously, the limited sample size did not allow us to isolate the impact of specific changes on the estimates. However, we still had to make decisions regarding the 2015 survey design based on the information that was available. This section describes the different sources of information that we had at our disposal and provides examples of how we used them.

The sources of information that we considered included the following:

- The NSDUH data that had already been collected;
- External data sources;
- Timing data;
- Interviewer debriefing items;
- Prior work in the field;
- Rates of missing data (i.e., "don't know" or "refused");
- Prior work within the survey; and
- "Other, specify" responses.

Using NSDUH data that had already been collected: Because the NSDUH 2012 field test sample was not large enough to accommodate one or more split samples, we decided to use the corresponding main survey data that had been collected in two different time periods as comparison datasets: main survey data from all four quarters of 2011 and main survey data that had been collected during a period in 2012 that was similar to the data collection period for the field test. The 2011 data were used as a comparison group because the large sample size and data collection over an entire year would improve the precision of estimates and help identify any seasonality in the estimates. The 2012 main survey data were used as a comparison dataset because the data were collected during roughly the same time period as the field test and therefore would not be subject to annual changes in prevalence. To further improve comparability with the field test data, main survey interviews that were completed in Alaska or Hawaii and interviews that were completed in Spanish were excluded from the comparison datasets. Thus, even though the essential survey conditions were not identical between the field test and the NSDUH main sample, we took

advantage of the continuous field data collection for the NSDUH to create reasonable comparison datasets.

Comparison with external data sources: Items that were collected for the first time in the field test had no point of reference from the current NSDUH survey for comparison purposes. Therefore, for these variables (e.g., height and weight, cell phone coverage), we turned to other large scale federal surveys to see how our field test estimates compared with estimates from these other surveys. For example, for the height and weight data, we examined estimates from the National Health Interview Survey (NHIS) and data from the National Health and Nutrition Examination Survey (NHANES), including respondent self-reports and physical measures. Based on these results, we were able both to verify that the data we collected appeared to be consistent with data from other sources and also to decide on the kinds of edit checks that we wanted to implement in the 2015 survey. However, one caution when comparing estimates with those from external sources is that methodological differences across the data sources can contribute to the results, regardless of whether the estimates are similar or different. Factors that can influence estimates include when the data are collected, the population surveyed, sample size, data collection language, survey mode, the actual questions, and question context.

Using timing data: As part of the field test, we collected timing data that told us not only how long the entire survey took to complete, but also what the completion times were for each of the modules. The timing data allowed us to evaluate if the survey was taking more or less time compared to our existing version. We also used the timing data to see if the difference in interview timings between persons based on education levels and age remained constant or changed with the partially redesigned questionnaire compared to the current questionnaire. For example, the increase in focus on past year prescription drug use and misuse in the field test could increase the difference in average times between younger and older respondents relative to the age group differences in interview times in the current survey. These changes to the prescription drug questions in the field test also could increase the number of interviews with extreme high interview times (i.e., outliers) relative to the number of outliers in the main survey.

Using interviewer debriefing items: We also had a set of interviewer debriefing items that the interviewers had to complete after the data were collected but before they could close out a case. The debriefing items included questions such as the following:

- Did the respondent remember receiving the lead letter?
- What comments, if any, did the respondent [R] make about the lead letter or in response to the lead letter?
- When did you give the respondent (or parent/guardian of youth respondent) the Q&A [question and answer] brochure?
- What comments, if any, did the respondent [R] (or parent/guardian) make about the Q&A [question and answer] brochure?
- Did the respondent make any comments about the interview being too long?
- Did the respondent have any questions or comments about the prescription drug questions in the ACASI [audio computer-assisted self-interviewing] section of the questionnaire?
- Please describe the respondent's [R's] comments about the prescription drug questions.
- Did the respondent have any questions or comments about the on-screen calendars in the ACASI [audio computer-assisted self-interviewing] section of the questionnaire? If the respondent asked how to access the calendar at any time during the ACASI portion of the interview, select "YES."
- What comments did the respondent [R] make about the on-screen calendars?
- Was a proxy used for the income and health insurance questions?
- Did the respondent have any questions or concerns about his/her answers being revealed to the proxy?
- Were there any problems with the proxy's understanding of the ACASI [audio computer-assisted self-interviewing] tutorial? What were they?

We used the three sources of information mentioned previously for the evaluation of the new prescription drug module. For our revised prescription drug module, we were interested not only in what the estimates would look like, but what the respondent burden would be. Therefore, we examined estimates, timing data, and interviewer debriefing items. The comparison of estimates indicated, as expected, that the change in the focus of the questions from lifetime to past year misuse in the field test, decreased the estimates of lifetime misuse and increased the estimates of past year use misuse in the field test relative to estimates in the NSDUH comparison datasets . The timing data indicated that older adults taking longer overall on the survey and the prescription drug sections in the field test than in the main survey. The timing data also indicated that the difference in timings between the older and

younger adults had increased, and the number of outlier cases with longer overall times was higher in the field test than in the main survey. Consistent with the effect of extreme values on means, however, the overall median survey times had not increased for the field test. The interview debriefing items showed that there were no large changes in how the respondents were reacting to the survey or the prescription drug module.

Using existing external research: As part of our attempts to decrease interviewer burden and increase respondent confidentiality, we moved some items related to employment, health insurance, income, and education from CAPI to ACASI administration. We found that there were notable differences in estimates of sources of income and in health insurance coverage between the NSDUH field test and both the NSDUH comparison datasets. This was of concern because these variables have become important for policy research related to the ACA. We could justify moving these items to ACASI in 2015 if the estimates based on ACASI data from the field test had either remained comparable to those from the main survey or we could demonstrate that ACASI administration improved the estimates. Improvements in the estimates based on ACASI administration could be demonstrated in one of two ways: (1) by showing that there was a theoretical basis for expecting these types of estimates to improve when the questions are self-administered; or (2) by comparing them with a 'gold standard' estimate. Thus, our first step was to do a literature review to examine whether other research had found that self-administration, such as through ACASI, improved the data quality for these specific measures compared with the data that were obtained through interviewer administration modes such as CAPI. We were unable to find anything that addressed this topic for these types of variables. We also compared estimates from our field test and comparison datasets to those from other sources such as NHIS, the American Community Survey, and the Current Population Survey. All of our comparisons with other sources of data showed that the estimates based on ACASI also did not agree with estimates from these other sources. This finding did not necessarily mean that the field test estimates were inaccurate. However, we did not have the evidence that we would have needed to justify a change in administration mode for these policy-relevant questions. We did have an option to wait and see if the same results were reproduced during the dress rehearsal that was scheduled for the fall of 2013. However, the dress rehearsal could have produced results that were not fully conclusive, similar to the QFT results, and there were concerns with waiting to make the decision. Therefore we made the decision to move these items back to the CAPI section.

Examining rates of missing data: If NSDUH respondents are routed to a question, item nonresponse (i.e., missing data) can occur if (1) respondents do not know an answer (e.g., they do not remember how old they were when they first used a substance); or (2) they exercise their right to refuse to answer the question. On the NSDUH, the rates of item nonresponse typically are fairly low. Therefore, any change in item nonresponse or having higher than expected levels of item nonresponse in new items may indicate that (1) there is a programming issue in the CAI instrument; (2) there are question wording or other issues with the question that are adversely affecting the cognitive tasks associated with answering the question; or (3) the question wording or subject matter are too "loaded" or sensitive for respondents. For example, a new question on family members serving in the military had a weighted missing data rate of 8.9 percent in the field test. This rate is very high by NSDUH standards. This appeared to be because the respondents were having difficulty knowing how to categorize certain family members. Therefore, for the dress rehearsal, the item was revised so that a definition of immediate family was directly added to the question. In the field test, the definition of immediate family was available to respondents as "help" text that they would not see unless they specifically requested this information. In addition, a new response category for "another member of the immediate family" was added for the dress rehearsal, and an "other, specify" follow-up question was added to allow respondents to specify their relationship to this other member of their immediate family. Based on preliminary results from the dress rehearsal, these changes appear to be reducing the item nonresponse rate for this question.

Generally higher rates of item nonresponse for certain items in the field test also were observed for health insurance and income items that were moved from CAPI to ACASI. This was a further indication that moving these items to ACASI did not improve the quality of the resulting data.

Using prior NSDUH methodological work and "other specify" responses

We also used existing methodological research to make decisions. For example, the decision on the imputation methods to use for the field test were driven by the small sample size (and therefore paucity of donors) and time. However, this decision could be made with relative confidence based on an earlier study that compared NSDUH estimates based on different imputation methods. We also examined 'other, specify' data to see whether respondents were selecting the 'other category' at rates that were higher than expected and whether the 'other specify' responses included those that were already available in the precoded response options. Either of these results would indicate

that (1) the overall question was not clear; (2) respondents were having difficulty choosing a response based on the wordings of the available response options (e.g., difficulty understanding the wordings, response options that appeared to overlap); or (3) a choice that is applicable or important to many respondents was missing from the available response options. In the first two of these situations, it could be less burdensome for respondents to choose "other" and then to type their response in the "other, specify" answer field than to try to determine which of the available response options best applied to them. In the third situation, the only way for respondents to answer the question accurately would be to choose "other" and then to specify their response.

Limitations and Conclusions

Despite our attempts to maximize all of the resources available to us, the small sample size for the field test (relatively speaking), affected our confidence in drawing certain conclusions from the data. For example, a lack of statistically significant differences between estimates in the field test and comparison datasets from the main survey does not allow us to be certain that there will be no effect in 2015 on the substance use estimates whose trend data we are trying to protect (e.g., estimates for alcohol, tobacco, marijuana and cocaine). For certain individual prescription drugs or groups of prescription drugs with a common active ingredient, low estimates of misuse in the field test (or no respondents who reported misuse) may be indicative of the small sample size rather than the actual prevalence in the population. Consequently, we may not adequately know the importance of certain prescription drugs for estimating misuse within an overall psychotherapeutic category (e.g., pain relievers) in advance of fielding the redesigned prescription drug module in the 2015 survey. The available sample size from the field test also may not allow us to adequately gauge the effects of changes to the prescription drug module on data in subsequent sections of the questionnaire that draw upon respondents' prior answers to the prescription drug items (e.g., substance dependence and abuse, and substance treatment). In addition, because of the multiple changes in the field test (e.g., new contact materials, new equipment, changes to the questionnaire), we are limited in our ability—especially within the constraints of time and resources—to draw firm conclusions about the most important reason for an observed difference, or the relative contributions of various changes to a result.

In particular, one of the challenges in effectively using all the data that are collected from a field test is making sure that there is sufficient time to process, analyze, and draw conclusions from the data. Sufficient time also is needed to implement any modifications based on the results. A field test becomes less valuable if all the information that could be gleaned from it cannot be learned in time for the next level of decision making. As with any complicated analyses, there generally are additional analyses that need to be conducted or tables that need to be re-specified, even if detailed analysis specifications were created ahead of time. There also are often unanticipated challenges. For example, we had not hypothesized that there would be the kinds of differences in ACASI and CAPI estimates that were observed in the field test. As part of our examination of why we observed these differences, we wanted to know if these differences were being driven by differences in response rates and potential differences in nonresponse bias between field test and comparison data that were not accounted for by the weighting procedures. However, there was not sufficient time to make this determination. There also was not sufficient time to evaluate the data before a decision was needed about whether to keep the health insurance and income items in ACASI for the dress rehearsal or to move them back to CAPI. If these items had been moved back to CAPI for the dress rehearsal, we could have evaluated the comparability of these items between the dress rehearsal and main survey data without the difference in mode of administration.

A field test is one of the many resources we can use to anticipate to the best of our ability what the best way to collect data appears to be and what the expected outcomes might be. Once a redesign is implemented, there is the post implementation of the estimates and quality of data to consider as well.